Automatic classification of LCA data

Shanlin Chen
chen.shanlin@hotmail.com
Instituto Superior Técnico, Universidade de Lisboa, Portugal

June 2020

Abstract

Life Cycle Assessment (LCA) is a useful tool for environment impact assessment and decision making in energy systems, meanwhile it's also complicated as it requires detailed information, time and computing resources. Many studies have been conducted to make LCA simpler in the categories of impact assessment like Cumulative Fossil Energy Demand, Carbon Footprints and Representativeness Index. Life Cycle Inventory (LCI), as an important part of a LCA study compiles all necessary information. This study aims to use machine learning algorithms to explore the data structures and hidden patterns in ecoinvent and exiobase LCI database, estimate the LCI of wind turbines with limited information for LCA.

Clustering results from different machine learning algorithms show that all the datasets of ecoinvent or exiobase seem to be one cluster with some random outliers by using inventories as features, location is a good feature to change the data structure but more information needed for better clustering.

It's possible to predict the LCI of a wind turbine from partial information through machine learning algorithms and mathematical methods, it's also practical to predict the total electricity production of wind turbines for better environmental impact assessment. Although the estimations of input materials and predicted amount of electricity may differ from the actual values, they are still good references for impact assessment as well as decision making.

Key-words: Life Cycle Assessment, Life Cycle Inventory, machine learning, wind turbine

1. Introduction

Life cycle assessment (LCA) is a technique compiling an inventory of relevant inputs and outputs of a product system, evaluating the potential environmental impacts associated with those inputs and outputs, and interpreting the results of the inventory and impact phases in relation to the objectives of the study.

After definition of the goal and scope, inventory analysis is the LCA step involves the compilation and

quantification of inputs and outputs for a given product system throughout its life cycle in the system boundary. Life cycle inventory (LCI) includes the collected data and its compilation from inventory analysis.

Life cycle impact assessment (LCIA) identifies and evaluates the amount and importance of the potential environmental impacts resulting from the LCI based on different LCIA methods. The inputs and outputs will be assigned to impact categories and their potential environmental impacts are quantified from characterization factors.

Then the results from impact assessment can be utilized to evaluate systems for improvements or decision-making with regards to environment and sustainability.

LCA is a very complex tool requires a large amount of detailed information and also time. Many studies have been conducted to offer possibilities to make LCA easier for implementation and interpretation but more on the basis of impact categories. Cumulative fossil energy demand might be a useful indicator for environment performance, because burning of fossil fuels is a major contributor to many environmental problems, it could be a screen indicator for energy production, material production, transport, global warming and resource depletion, but not for waste treatment and land use [1]. Representativeness index aims to reduce the size of environment impact categories by correlation analysis, contribute to the interpretation of LCA results through pointing to specificities of inventories and identifying the main representative impact categories [2]. Resource footprint and damage footprint are good proxies of environmental damage, while resource footprints accounted for more than 90% of the variation in the damage footprints [3]. And studies aimed at comparing different LCIA methods generally find they are highly correlated between impact categories from different methods [3,4], so it's not very influential to choose the impact assessment methods and using representatives for other LCIA methods can simplify LCA with less calculation based on the strong correlations between the impact categories.

LCI is compiled of all necessary information for LCA. Ecoinvent [5] and exiobase [6] are LCI databases providing background data for LCA. The ecoinvent data is a widely used LCI database offers full information for over 18000 datasets and 3000 products in up to 140 countries. Ecoinvent datasets cover all relevant environmental flows, including resource extraction, land use, emissions as well as materials and energy supplies in a global or regional level. Exiobase consists of global Multi-Regional Environmentally Extended Supply-Use (MR-SUT) table and Input-Output (MR-IOT) table. A large

number of countries are involved via MR-SUT and MR-IOTs for estimating emissions and resource extractions by industry, as well as analysis of environmental impacts related with the final consumption of products.

There are a lot of detailed information in LCI databases, like electricity production, transport, they are split into groups based on engineering judgments. However, we can use machine learning algorithms to create groups based on their measured attributes, including LCI, LCA indicators. Using data-based classification could reduce uncertainty and better help us understand when to or not to split datasets into different groups, which should in turn help us understand the data better.

Exploring the data patterns in LCI databases is the first part of this thesis work, which aims to understand and use data in a more efficient way. Applying machine learning algorithms in creating LCI is the second part, which gives a possibility to perform LCA with even limited information.

2. Methods

Activity clustering aims to find the hidden patterns or structures for activities from LCI database by unsupervised machine learning, with a purpose of better understanding and more efficient utilization.

2.1 Unsupervised machine learning

In pattern recognition problems, the training data is a set of input vector without any corresponding target values are known as unsupervised learning. Unsupervised machine learning can be used for density estimation to determine the distribution of data within the input space, or for visualization by projecting data from a high-dimensional space down to two or three dimensions, or to discover groups of similar samples within the data, which is also called clustering [7].

There are a lot of different algorithms for clustering by using distance, density or similar measurements between data points to identify patterns.

Partitional clustering are methods used to classify samples within a dataset into groups based on their similarity, like k-means [8] algorithm, in which each cluster is represented by the center or means of data points in the cluster, so the k-means method is sensitive to outliers. And partitional clustering methods require the number of clusters before clustering, however, this is rarely the case [9], because the purpose for pattern recognition problems is to find the hidden information, and the number of clusters is a part of this information to be found.

Density-based algorithms consider that clusters are based on connectivity and density functions, like DBSCAN [10], in which the clusters are condensed and separated by areas of low density. Unlike k-means, the clusters found by DBSCAN can be any shape. There are two important parameters to this method, one is the maximum radius of the neighborhood and the other is minimum number of points within the radius of a point. Therefore, DBSCAN can handle outliers but is sensitive to parameters.

Hierarchical clustering is generally a number of family clustering algorithms which find clusters by merging small clusters into larger ones or splitting large clusters. The dendrogram is a hierarchy of clusters and shows how clusters are related. There is no outlier in hierarchical clustering, one cluster can have only one sample. Agglomerative Clustering is a bottom-up approach, which merges small clusters into larger ones based on different linkage criteria [9].

Meanwhile, there are also other clustering methods, OPTICS [11] has many similarities with DBSCAN, while OPTICS builds a reachability graph that change the maximum radius from a single value to a value range. Birch [12] deals with large datasets by compressing the original data into a set of clustering feature nodes, and then clustering the subset instead of initial dataset to reduce the memory requirement. Mean Shift [13] is a centroid based algorithm, which works by updating centroids to be the mean of points in a given region like k-means, but the number of clusters prior clustering is not required.

LCI estimation gives a possibility to perform LCA studies with even limited information. Estimating the LCI of wind turbines is the specific example in this study by supervised machine learning algorithms and some other mathematical methods.

There is usually large amount of data involved in machine learning, so one challenge is gathering the data for training a model and testing. Fortunately, datasets about some dimensional parameters do exist, like rotor diameter, hub height, rated power of wind turbines, therefore, machine learning could be helpful when just using partial information to generate the life cycle inventory of a wind turbine. Machine learning can also be applied in predicting service time and total electricity production, which gives information to calculate the impact in the same unit for better comparison with other projects.

In order to train a machine learning model for prediction, several steps are usually involved. Data collection is the first, and then it comes to data preparation, which is also known as data preprocessing, it requires data visualization to get insight in data, outlier detection, dimensionality reduction in case of high-dimensional problems, and feature scaling to ensure the data can be easily interpreted, next step

is modeling with different machine learning algorithms and model evaluation by different performance indexes, if the trained models were not good enough, feature extraction and selection would be applied to create more features, modeling is more likely a iterative step, which requires model evaluation and selection. Finally, it's model interpretation, which aims to apply learned patterns or information for predictions.

2.2 Supervised machine learning

Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as supervised learning problems [7]. When the desired output variable is continuous, then it is called a regression problem. There are many regression methods in machine learning can be applied to train the model and then estimate the some unknown values.

Linear regression typically involves a linear combination of input parameters, and it can be presented by

$$\mathrm{y}(\mathrm{x}, \mathrm{w}) = w_0 + w_1 x_1 + \cdots + w_p x_p \quad (1)$$

Where x=(x_1, x_2, …, x_p) is input variable, and w=(w_1, w_2, …, w_p) is the vector of coefficients.

There are limitations of simple linear models, therefore we extend the models by polynomial processing to generate polynomials as inputs or by linear combinations of nonlinear functions of input variables.

Decision trees are a non-parametric supervised learning method, aims to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features[14]. Decision trees usually learn from data to approximate a fitting curve with a set of if-then-else decision rules. Deeper tree has more complex decision rules and fits the model better. Random forest regression is an averaging algorithm based on decision trees.

Neural network, which is also known as Multi-layer Perceptron, is a supervised learning algorithm used for both classification and regression. The input layer consists of a set of neurons known as input features. Every neuron in the hidden layer transforms the values from previous layer with a weighted linear summation, followed by an activation function, which is also called sigmoid function. The output layer gets the values and transforms them into outputs [14].

Support vector machine (SVM) [15] is considered by many to be the most powerful 'black box' learning algorithm in classification, regression and novelty detection, but more for classification problems, since it looks for a hyper-plane or set of hyper-planes to separate samples.

3. Results and discussion

3.1 LCI database clustering

The curse of dimensionality stands for various phenomena that cause difficulties when analyzing or processing data in high-dimensional spaces, for a clustering problem, the dimensionality curse exists as well, fortunately many techniques have been developed to reduce the dimensions of the input dataset. When setting the inventories as input features for LCI database clustering, it's definitely a high-dimensional clustering case as the number of features exceed a thousand, mathematical methods like PCA and correlation analysis are used to reduce the dimensions and retain as much information as possible, besides, engineering solutions like combining similar elementary flows or removing less important flows can help as well.

With different dimensionality reduction techniques, input features as well as clustering algorithms being applied for ecoinvent and exiobase database, no hidden patterns or other useful information have been founded yet. The possible reasons might be most of the activities in ecoinvent or exiobase, no matter it's about materials production or providing services, they all related with energy supply, in other words, energy is the basic requirement for the activities, therefore, when comes to the exchanges with bioflows, they may have similar values. Meanwhile most exchanges are pretty small, even though there are also large numbers, they don't have enough influence to change the overall patterns.

It's worth mentioning that dimensionality reduction from an engineering point might also be useful apart from mathematical methods. Adding proper new features will definitely improve the clustering results but it is not easy to find appropriate ones. Many LCIA methods are highly correlated, so maybe it's a good way to implement LCA calculations for certain LCIA methods and then use them as references for other categories of impact assessment and decision making.

3.2 Estimating Life cycle inventory for wind turbines

Another issue concerning LCI is that the LCI requires very detailed information from cradle to grave, but in most case they are not available. Then here is another chance to use machine learning algorithms to predict missing data based on limited information, which is known as a supervised machine learning problem, labeled data are being used to train the model and then applied to estimate the unknown data with some initial information.

Different machine learning algorithms are applied for estimating the LCI of wind turbine and also for total electricity production. In most cases, multiple linear regression and random forest regression give better results, polynomial processing and feature extraction will create new features and improve models' performance, and feature selection reduces the dimensionality and simplifies the model. Random forest and neural network are particularly useful for samples with a few features, while support vector machines turn out to be over fitting as the training score is usually high but the validation score is low.

There is a tradeoff between bias and variances, to avoid under-fitting or over-fitting, the dataset is usually divided randomly into a training set and a smaller testing set. Model selection can be based on learning curves or some performance indexes. Generally, models with higher testing score and lower errors will be selected for prediction.

In the LCI estimation of wind turbines, a model combined machine learning algorithms and mathematical methods is developed and can be tailored based on the known information. The worst case is that only rated power and location (Onshore or Offshore) of a wind turbine are available, then the dimensions like rotor diameter, hub height, masses of rotor, nacelle, tower can be estimated by machine learning models, sizing of foundation, transformers and required cables are calculated based some mathematical methods and assumptions. Materials breakdown is based on split ratios from some detailed inventories of wind turbines with different capacities, transport requirement, maintenance and end-of-lie disposal strategy is taken from a LCA report.

The amount of all materials is proportional to the nominal power, which is the same tendency for actual and predicted values, however, the total mass of a wind turbine (including foundation) is somehow over estimated, this is mainly because of the over prediction of foundation, whose main materials are reinforcing steel and concrete, this is why the predicted values of steel and concrete are larger than actual ones, so the model to get the mass of foundation should be modified based on more data. Fiberglass is over predicted, the reason for this might be that the assumption regarding materials breakdown for rotor is very different from these examples. Fiberglass mainly comes from rotor and nacelle, and there is a compromise between fiberglass and cast iron, overestimation of fiberglass leads to decrease of cast iron. Cables for electricity transmission is assumed made of Copper and polymer insulation layers, but actually both Copper and Aluminum (more for economical reason and lower

density) are used as conductors, this is why there is a under estimation of Aluminum, another factor is the transmission distance, which strongly depends on the layout of wind farms.

It is also important to notice that even wind turbines with the same manufacturer and capacity may have different contents of materials [17,19,20,21], which makes the estimated life cycle inventory less accurate.

When it comes to impact assessment, apart from the gaps of input materials caused by estimation, there are some others issues, first, life time of wind turbine and total electricity production, in the reports [16-19], the life time is 20 years and the electricity production is calculated based an annual average load factor over 40%, which may lead to the total electricity production is much more than the actual and predicted values, thus the indexes like GWP/kWh might be smaller. Second, the background processes, since the location of wind farm is either barely mentioned or in a more general way, it's not possible to adjust the background processes to match the location, therefore, datasets from a global level is used as first choices, which may also have negative effect on the impact assessment.

After estimating the LCI of a wind turbine, it's possible to perform LCA to see the environmental impacts, however, for a better comparison with other projects, it's necessary to know the service time and lifetime electricity production, with available information from Danish Ministry of Energy, it's possible to train models to forecast the life time and total electricity production of wind turbines. Although the model for service time doesn't give satisfying result, it's still better than just setting a life time of 20 years. Furthermore, predicting total electricity without life time is still feasible with a lower accuracy.

With more available data, it's achievable to build a model for estimating the entire LCI of a wind turbine by machine learning algorithms. Although the estimated LCI may not be highly accurate, it would give more support for environmental impact assessment and provide references for better decision making especially when the initial information is limited. And this methodology can also be applied for other renewable energy systems like photovoltaic panels and geothermal power plants.

Acknowledgement

The author would like to express his gratitude to all the persons who made contributions to the work described in the study. Additionally, the author would like to thank Dr. Christopher Lucien Mutel and Prof. Susana Margarida da Silva Vieira for their great supervision and support and Dr. Romain Sacchi for providing information and remarkable insights to bring machine learning into Life Cycle Assessment

studies. Finally, the author wishes to thank Paul Scherrer Institut, InnoEnergy, SUT and IST for all the materials, resources and thoughtful supports received.

References

[1] Mark A. J. Huijbregts, Linda J. A. Rombouts, Stefanie Hellweg, Rolf Frischknecht, A. Jan Hendriks, Dik van de Meent, Ad M. J. Ragas, Lucas Reijnders, and Jaap Struijs. Is cumulative fossil energy demand a useful indicator for the environmental performance of products? Environmental Science & Technology (2006): 641-648.

[2] Esnouf A, Heijungs R, Coste G, Latrille É, Steyer J.P, and Hélias A. A tool to guide the selection of impact categories for LCA studies by using the representativeness index. Science of the Total Environment, 2019, 658: 768-776.

[3] Steinmann Z J N, Schipper A M, Hauck M, Giljum S, Wernet G, and Huijbregts M.A. Resource footprints are good proxies of environmental damage. Environmental science & technology, 2017, 51(11): 6360-6366.

[4] Berger M, Finkbeiner M. Correlation analysis of life cycle impact assessment indicators measuring resource use. The International Journal of Life Cycle Assessment, 2011, 16(1): 74-81.

[5] Wernet G, Bauer C, Steubing B, Reinhard J, Moreno-Ruiz E, and Weidema B. The ECOINVENT database version 3 (part I): overview and methodology. The International Journal of Life Cycle Assessment, 2016, 21(9): 1218-1230.

[6] Merciai, S. and J. Schmidt. Methodology for the Construction of Global Multi-Regional Hybrid Supply and Use Tables for the EXIOBASE v3 Database. Journal of Industrial Ecology, 2018, 22(3): 516-531.

[7] Bishop C M. Pattern recognition and machine learning[M]. springer, 2006.

[8] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding[R]. Stanford, 2006.

[9] Grira N, Crucianu M, Boujemaa N. Unsupervised and semi-supervised clustering: a brief survey[J]. A review of machine learning techniques for processing multimedia content, 2004, 1: 9-16.

[10] Ester M, Kriegel H P, Sander J, and Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996, 96(34): 226-231.

[11] Ankerst M, Breunig M M, Kriegel H P, Ng R.T, and Sander J. OPTICS: ordering points to identify the clustering structure. ACM Sigmod record, 1999, 28(2): 49-60.

[12] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large

databases. ACM Sigmod Record, 1996, 25(2): 103-114.

[13] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on pattern analysis and machine intelligence, 2002, 24(5): 603-619.

[14] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, and Vanderplas J. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 2011, 12: 2825-2830.

[15] Evgeniou T, Pontil M. Support vector machines: Theory and applications. Advanced Course on Artificial Intelligence. Springer, Berlin, Heidelberg, 1999: 249-257.

[16] Life Cycle Assessment of Electricity Production from an onshore power plant based on Vestas V82-1.65 MW turbines. Vestas Wind Systems A/S, 2006.

[17] Razdan P, Garrett P. Life Cycle Assessment of Electricity Production from an onshore V110-2.0 MW Wind Plant. Vestas Wind Systems A/S, 2015.

[18] Garrett P, Ronde K. Life Cycle Assessment of Electricity Production from an onshore V90-3.0 MW Wind Plant. Vestas Wind Systems A/S, 2013.

[19] Razdan P, Garrett P. Life Cycle Assessment of Electricity Production from an onshore V112-3.45 MW Wind Plant. Vestas Wind Systems A/S, 2017.

[20] Garrett P, Ronde K. Life Cycle Assessment of Electricity Production from a V80-2.0 MW Gridstreamer Wind Plant. Vestas Wind Systems A/S, 2011.

[21] Razdan P, Garrett P. Life Cycle Assessment of Electricity Production from an onshore V105-3.45 MW Wind Plant. Vestas Wind Systems A/S, 2017.